Rhythmic unit extraction and modelling for automatic language identification

Jean-Luc Rouas, Jérôme Farinas, François Pellegrino, Régine André-Obrecht

Title: Rhythmic Unit Extraction and Modelling for Automatic Language Identification

Authors : Jean-Luc Rouas[1], Jérôme Farinas[1], François Pellegrino[2], Régine André-Obrecht[1]

[1] Institut de Recherche en Informatique de Toulouse UMR 5505 CNRS – Institut National Polytechique de Toulouse – Université Paul Sabatier – Université Toulouse 1, France

[2] Laboratoire Dynamique Du Langage UMR 5596 CNRS – Université Lumière Lyon 2, France

{rouas@irit.fr, Jerome.Farinas@irit.fr, Francois.Pellegrino@univ-lyon2.fr, obrecht@irit.fr}

Corresponding author :
François PELLEGRINO (Francois.Pellegrino@univ-lyon2.fr)
Postal Address : DDL – ISH
14, avenue Berthelot
69363 LYON CEDEX 7
FRANCE
Tel: +33 4 72 72 64 77
Fax: +33 4 72 72 65 90

ABSTRACT

This paper deals with an approach to Automatic Language Identification based on rhythmic modelling. Beside phonetics and phonotactics, rhythm is actually one of the most promising features to be considered for language identification, even if its extraction and modelling are not a straightforward issue. Actually, one of the main problems to address is *what* to model. In this paper, an algorithm of rhythm extraction is described: using a vowel detection algorithm, rhythmic units related to syllables are segmented. Several parameters are extracted (consonantal and vowel duration, cluster complexity) and modelled with a Gaussian Mixture. Experiments are performed on read speech for 7 languages (English, French, German, Italian, Japanese, Mandarin and Spanish) and results reach up to 86 ± 6% of correct discrimination between stress-timed mora-timed and syllable-timed classes of languages, and to 67 ± 8% percent of correct language identification on average for the 7 languages with utterances of 21 seconds. These results are commented and compared with those obtained with a standard acoustic Gaussian mixture modelling approach (88 ± 5% of correct identification for the 7-languages identification task).

KEYWORDS

Rhythm modelling; Language identification; Rhythm typology; Asian Languages; European languages

1. INTRODUCTION

Automatic Language Identification (ALI) has been studied for almost thirty years, but the first competitive systems appeared during the 90s. This recent attention is related to 1. the need for Human-Computer Interfaces and 2. the remarkable expansion of multilingual exchanges. Indeed, in the so-called information society, the stakes of ALI are numerous, both for multilingual Human-Computer Interfaces (Interactive Information Terminal, Speech dictation, etc.) and for Computer-Assisted Communication (Emergency Service, Phone routing services, etc.). Moreover, accessing the overwhelming amount of numeric audio (or multimedia) data available may take advantage from content-based indexing that may include information about the speakers' languages or dialects. Besides, linguistic issues may also be addressed: the notion of linguistic distance has been implicitly present in linguistics typology for almost a century. However, it is still difficult to define, and ALI systems may shed a different light on this notion since correlating automatic, perceptual and linguistic distances may lead to a renewal of the typologies and to a better understanding of the close notions of languages and dialects.

At present, state-of-the-art approaches consider phonetic models as *front-end* providing sequences of discrete phonetic units decoded later in the system, according to language-specific statistical grammars (see Zissman & Berkling, (2001) for a review). The recent NIST 2003 Language Recognition Evaluation (Martin & Przybocki, 2003) has confirmed that this approach is quite effective since the error rate obtained on a language verification task using a set of twelve languages is under 3% for 30-second utterances (Gauvain et al., 2004). However, other systems modelling global acoustic properties of the languages are also very efficient, and yield about 5% error on the same task (Singer et al., 2003). These systems, that take advantage either of speech or speaker recognition techniques, perform quite well. Still, very few systems are

trying to use other approaches (e.g. prosodics) and results are much poorer than those obtained with the phonetic approach (for example the combination of the standard OGI "temporal dynamics" system based on a *n*-gram modelling of sequences of segments labelled according their Fo and energy curves yields about 15-20% of equal error rate with three languages of the NIST 2003 campaign task and corpus (Adami & Hermanski, 2003)). However, these alternative approaches may lead to improvements, in terms of robustness in noisy conditions, number of languages recognized or linguistic typology. Further research efforts have to be made to overcome the limitations and to assess the contributions of those alternative approaches.

The motivations of this work are given in Section 2. One of the most important is that prosodic features carry a substantial part of the language identity that may be sufficient for humans to perceptually identify some languages (see Section 2.2). Among these supra-segmental features, rhythm is very promising both for linguistic and automatic processing purposes (Section 2). However, coping with rhythm is a tricky issue, both in terms of theoretical definition and automatic processing (Section 3). For these reasons, the few previous experiments which aimed at language recognition using rhythm were based on hand-labelled data and/or have involved only tasks of language discrimination[1] (Thymé-Gobbel & Hutchins, 1999; Dominey & Ramus, 2000). This paper addresses the issue of automatic rhythm modelling with an approach that requires no phonetically labelled data (Section 4). Using a vowel detection algorithm, rhythmic units somewhat similar to syllables and called *pseudo-syllables* are segmented. For each unit, several parameters are extracted (consonantal and vowel duration, cluster complexity) and modelled with a Gaussian Mixture. This approach is applied to 7 languages (English, French, German, Italian Japanese, Mandarin and Spanish) using the MULTEXT corpus of read speech. Descriptive statistics on pseudo-syllables are computed and the relevancy of this modelling is

[1] *Language discrimination* refers to determining to which of two candidate languages L_1-L_2 an unknown utterance belongs to. *Language identification* denotes more complex tasks where the number of candidate languages is more than two.

assessed with two experiments aiming at 1. discriminating languages according to their rhythmic classes (stress-timed vs. mora-timed vs. syllable-timed) and 2. identifying the 7 languages. This rhythmic approach is then compared to a more standard acoustic approach (Section 5).

From a theoretical point of view, the proposed system focuses on the existence and the modelling of rhythmic units. This approach generates a type of segmentation that is closely related to a syllabic parsing of the utterances. It leaves aside other components of rhythm related to the sequences of rhythmic units or that span over whole utterances. These considerations are discussed in Section 6.

2. MOTIVATIONS

Rhythm is involved in many processes of the speech communication. Though it has been neglected for long, several considerations lead to reconsider its role both in understanding and production processes (Section 2.1), and especially in a language identification framework (Section 2.2). Moreover, researchers have tried to take rhythm into consideration for automatic processing purposes for a while, both in speech synthesis and recognition tasks, leading to several rhythm-oriented approaches (Section 2.3). All these considerations emphasize both the potential use of an efficient rhythm model and the difficulty to elaborate it. It leads us to focus on the possible use of rhythmic features for ALI (Sections 3 and 4).

2.1. Linguistic definition and functions of rhythm

Rhythm is a complex phenomenon that has long been said to be a consequence of other characteristics of speech (phonemes, syntax, intonation, etc.). However, an impressive amount of experiments tends to prove that its role may be much more than a mere side effect in the speech communicative process.

According to the Frame/Content theory (MacNeilage, 1998; MacNeilage & Davis, 2000), speech production is based on superimposing a segmental content into a cyclical frame. From an

evolutionary point of view, this cycle probably evolved from the ingestive mechanical cycles shared by mammals (e.g. chewing) via intermediate states including visuofacial communication controlled at least by a mandibular movement (lipsmacks, etc.). Moreover, the authors shed the light on the status of the syllable both as an interface between segments and suprasegmentals and as the *frame*, a central concept in their theory: convoluting the mandibular cycle with a basic voicing production mechanism results in a sequence of CV syllables composed of a closure and a neutral vowel. Additional experiments on serial ordering errors made by adults or children (e.g. Fromkin, 1973; Berg, 1992) and child babbling (MacNeilage et al. 2000; Kern et al., to appear) are also compatible with the idea that the mandibular oscillation provides a rhythmic baseline in which segments accurately controlled by articulators take place.

A huge amount of psycholinguistics studies also draw attention to the importance of the rhythmic units in the complex process of language comprehension. Most of them consider that a rhythmic unit – roughly corresponding to the syllable combined with an optional stress pattern – plays an important role as an intermediate level of perception between the acoustic signal and the word level. The exact role of these syllables or syllable-sized units has still to be clearly identified: whether the important feature is the unit itself (as a recoding unit) or its boundaries (as milestones for the segmentation process) is still in debate. The ones claim that the syllable is the main unit in which the phonetic recoding is performed before lexical access (Mehler et al., 1981). The others propose an alternative hypothesis in which syllables and/or stress provide milestones to parse the acoustic signal into chunks that are correctly aligned with the lexical units (Cutler & Norris, 1988). In this last framework, the boundaries are more salient than the content itself, and no additional hypothesis is made on the size of the units actually used for lexical mapping. Furthermore, recent experiments point out that the main process may consist in locating the onset rather than raw boundary detection (Content et al., 2000; Content et al., 2001).

These studies show that rhythm plays a key role in the speech communication process. Similarly, several complementary aspects could have been mentioned but they are beyond the

scope of this paper[2]. However, several questions regarding the nature of the rhythm phenomenon are still open. First of all and as far as the authors know, an uncontroversial definition of rhythm does not exist yet even if most researchers may agree on the notion that speech rhythm is related to the existence of a detectable phenomenon that occurs evenly in speech. Crystal proposes to precisely define rhythm as "the regular perception of prominent units in speech" (Crystal, 1990). We prefer not to use the concepts of *perception* and *unit* because they narrow the rhythmic phenomenon with *a priori* hypotheses: according to Crystal's definition, rhythm can be considered as the alternation of prominent units with less prominent ones, but defining those units is far from straightforward; The alternation of stressed/unstressed syllables results in one kind of rhythm, but the voiced/unvoiced sound sequences may produce another type of rhythm, and so do consonant/vowel alternations or short/long sound sequences, etc. Moreover, rhythm may arise from the even occurrence of punctual *events* and not *units* (like beats superimposed on other instruments in music).

Another question concerns the actual role of the *syllable*. Whether it is a cognitive unit or not is still in debate. Though, several experiments and measures indicate that syllables or syllable-sized units are remarkably salient and may exhibit specific acoustic characteristics. Since the early 1970s, several experiments have indicated that the human auditory system is especially sensitive to time intervals spanning from 150 to 300 ms clearly compatible with average syllable duration[3]. These experiments, based on various protocols (forward and backward masking effect, ear switching speech, shadowing repetition, etc.) showed that this duration

---

[2] See Levelt (1994) for his model of speech production; see also Boysson-Bardies et al., 1992; Mehler et al., 1996; Weissenborn & Höhle, 2001; Nazzi & Ramus, 2003 for the role of rhythm in early acquisition of language.

[3] Greenberg (1998) reports a mean duration of 200 ms for spontaneous discourse on the Switchboard English database.

roughly corresponds to the size of a human perceptual buffer (see for example Massaro, 1972; Jeasted et al., 1982; O'Shaughnessy 1986). More recently, experiments performed with manipulated spectral envelopes of speech signals showed the salience of the modulation frequencies between 4 and 6 Hz in perception (Drullman et al., 1994). Hence, all these findings support the syllable as a relevant rhythmic unit. In addition, acoustic measurements made on a corpus of English spontaneous speech emphasize also its prominence (Greenberg, 1996 & 1998). This study showed that, as far as spectral characteristics are concerned, syllable onsets are in general less variable than nuclei or codas. It also highlights that co-articulation effects are much larger within each syllable than between syllables. Both effects result in the fact that syllable onsets vary less than other parts of the signal and consequently may provide at least reliable anchors for lexical decoding. Besides this search for the intrinsic *nature* of rhythm, perceptual studies may also improve our knowledge of its intrinsic *structure*. Using speech synthesis to simulate speech production, Zellner-Keller (2002) concluded that rhythm structure results from a kind of convolution of a temporal skeleton with several layers, from segmental to phrasal, in a complex manner that can be partially predicted.

One of the main conclusions is that temporal intervals ranging from 150 to 300 ms are involved in speech communication as a relevant level of processing. Moreover, many cues draw attention to this intermediate level between acoustic signal and high level tiers (syntax, lexicon). At this moment, it is not evident to assess if the relevant feature is actually a rhythmic *unit* by itself or a rhythmic *beat*. However syllable-sized units are salient from a perceptual point of view and may have acoustic correlates that facilitate their automatic extraction. Next section deals with the experimental assessment of these correlates in perceptive language identification tasks.

2.2. Rhythm and Perceptual Language Identification

Language identification is an uncommon task for many adult human speakers. It can be viewed as an entertaining activity by the most questioning ones but most adult human beings living in a monolingual country may consider that it is of no interest. However, the situation is quite

different in multilingual countries where numerous languages or dialects may be spoken on a narrow geographical area. Furthermore, perceptual language identification is an essential challenge for children who acquire language(s) in that kind of multilingual context: it is then utterly important for them to distinguish which language is spoken in order to acquire the right language-dependent phonology, syntax, and lexicon. During the last two decades, several experiments have investigated the efficiency of the human being as a language recognizer (see Barkat-Defradas et al., 2003, for a review). Three major types of features may help someone to identify a language: 1. Segmental features (the acoustic properties of phonemes and their frequency of occurrence), 2. Supra-segmental features (phonotactics, prosody), and 3. High level features (lexicon, morpho-syntax). The exact use made of each set of features is unclear yet and it may actually differ between newborn children and adults.

For example, several experiments have proved that newborns, as early as the very first days, are able to discriminate between their mother tongue and some foreign languages that exhibit differences at the supra-segmental level (see Ramus, 2002, for a review). Whether newborns take advantage from rhythm alone or from both rhythm and intonation is an open issue. It is likely that both levels provide cues that are weighted as function of the experimental conditions (languages, noise, and speech rate) and maybe according to individual strategies. Assessing these adult human capacities to identify foreign languages is a complex challenge since numerous parameters may influence this ability. Among them, the subject's mother tongue and his personal linguistic history seem to be key factors that prove difficult to quantify. Since the end of the 1960s, quite a few studies have tackled this question. Depending on whether they are implemented by automatic speech processing researchers or linguists, the purposes differ. The former intend to use these perceptual experiments as benchmarks for ALI systems, while the latter investigate the cognitive process of human perception. More recently, this kind of experiments has been viewed as a way to investigate the notion of *perceptual distance* among languages. In this framework, the aim is to evaluate the influence of the different levels of linguistic description in the cognitive judgment of language proximity.

From a general point of view all these experiments have shown the noteworthy capacity of human subjects to identify foreign languages after a short period of exposure. For example, one of the experiments reported by Muthusamy et al., (1992) indicates that native English subjects reach a score of 54.2% of correct answers when identifying 6-second excerpts pronounced in 9 foreign languages. Performances varied significantly from one language to another, ranging from 26.7% of recognition for Korean to 86.4% of recognition for Spanish. Additionally, subjects were asked to explain which cues they had considered to make their decision. Their answers revealed the use of segmental features (manner and place of articulation, presence of nasal vowels, etc.), supra-segmentals (rhythm, intonation, tones) and "lexical" cues (iteration of the same words or pseudo-words). However these experiments raise numerous questions about the factors influencing the recognition capacity of the subjects: the number of languages that they have been exposed to, the duration of the experimental training, etc. Following Muthusamy, several researchers have tried to quantify these effects. Stockmal, Bond and their colleagues (Stockmal et al., 1996; Stockmal et al., 2000; Bond & Stockmal, 2002) have investigated several socio-linguistic factors (geographical origin of the speakers, languages known by the subjects, etc.) and linguistic factors (especially rhythmic characteristics of languages). In a similar task based on the identification of Arabic dialects our group has shed light on the correlation between the structure of the vocalic system of the dialects and the perceptual distances estimated from the subjects' answers (Barkat et al., 2001). The results reported by Vasilescu et al., (2000) in an experiment of discrimination between romance languages may be interpreted in a similar way. Other studies focus on the salience of supra-segmentals in perceptual language identification. From the first experiments of Ohala and Gilbert, (1979) to the recent investigations of Ramus, using both natural and synthesized speech, they prove that listeners may rely on phonotactics, rhythm, and intonation patterns to distinguish or identify languages, even if segmental information is lacking.

Even if the cognitive process leading to language identification is multistream (from segmental acoustics to suprasegmentals and higher level cues), no model of integration has been derived

yet. Moreover, building such a model seems to be still out of range since even the individual mechanisms of perception at each level are still puzzling. At the segmental level, most researchers are working with reference to the motor theory of speech perception (Liberman & Mattingly, 1985) searching arguments that would either confirm or invalidate it. At the suprasegmental level, the perception of rhythm has been mainly studied from a musical point of view, even if comparisons between music and speech perception are also studied (e.g. Todd & Brown, 1994; Besson & Schön, 2001) and if technological applications (e.g. speech synthesis) have lead researchers to evaluate rhythm (see next section).

2.3. Rhythm and syllable-oriented automatic speech processing

Many studies aiming at taking advantage from rhythmic and prosodic features for automatic systems have been developed through the last decades and achieved most of the time disappointing results. Nevertheless several authors consider that this is a consequence of the difficulty to model suprasegmental information and put forward the major role of prosody and temporal aspects in speech communication processes (see for example Zellner Keller & Keller, 2001 for speech synthesis and Taylor et al., 1997 for speech recognition).

Beside its role in the parsing of sentence into words (Cutler & Norris, 1988; Cutler, 1996), prosody constitutes sometimes the only means to disambiguate sentences, and it often carries additional information (mood of the speaker, etc.). Even when focusing on the acoustic-phonetic decoding, suprasegmentals may be relevant at two levels: first of all, segmental and suprasegmental features are not independent, and thus, the suprasegmental level may help to disambiguate the segmental level (e.g., see the correlation between stress accent and pronunciation variation in American English (Greenberg et al., 2002)). Moreover, as it has been argued above, suprasegmentals and especially rhythm, may be a salient level of treatment as itself for humans and probably for computational models. Speech synthesis is an evident domain where perceptual experiments have shown the interest of syllable-length units for the

naturalness of synthesized speech (Keller & Zellner, 1997). Additionally, rhythm and rhythmic units may play a major role in Automatic Speech Recognition: from the proposal of the syllable as a unit for speech recognition (Fujimura, 1975) to the summer workshop on "Syllable Based Speech Recognition" sponsored by the Johns Hopkins University (Ganapathiraju, 1999), attempts to use rhythmic units in automatic speech recognition and understanding have been numerous. Disappointingly, most of them failed to improve the standard speech recognition approach based on context-dependent phone modelling (for a review, see Wu, 1998). However, the definitive conclusion is not that suprasegmentals are useless, but instead, that the phonemic level may not be the suitable time scale to integrate them and that larger scales may be more efficient. We have already mentioned that co-articulation effects are much greater *within* each syllable than *between* syllables in a given corpus of American English spontaneous speech (Greenberg, 1996). Context-dependent phones are well-known to efficiently handle this co-articulation. However, their training needs a big amount of data, and consequently they can not be used when few data are available (this happens especially in multilingual situations: the state-of-the-art systems of ALI are based on context-independent phones (Singer et al., 2003; Gauvain et al., 2004). Thus, syllable-sized models are a promising alternative with limited variability at the boundaries. However, several unsolved problems limit the performance of the current syllable-based recognition systems and the main problem may be that syllable boundaries are not easy to identify, especially in spontaneous speech (e.g. Content et al., 2000 for a discussion on ambisyllabicity and resyllabification). Thus, combining phoneme-oriented and syllable-oriented models in order to take several time scales into account may be a successful approach to overcome the specific limits of each scale (Wu, 1998). Finally, syllable-oriented studies are less common in the fields of speaker and language identification. Among them, we can however distinguish approaches adapted from standard phonetic or phonotactic approaches to syllable-sized units (Li, 1994 and more recently Antoine et al., 2004 for a "syllabotactic" approach and Nagarajan and Murthy, 2004 for a syllabic Hidden Markov Modelling) from those trying to model the underlying rhythmic structure (see section 4.1).

This section showed that 1. Rhythm is an important mechanism of speech communication involved in comprehension and production processes; 2. It is difficult to define, to handle, and most of all, to efficiently model; 3. Syllable or syllable-like units may play an important role in the structure of rhythm. Furthermore, experiments reported above clearly demonstrate that different languages may be different from the rhythmic perspective and that these differences may be perceived and used in a perceptual language identification task. Next section deals with these differences, both in terms of linguistic diversity and its underlying acoustic parameters.

3. THE RHYTHM TYPOLOGY AND ITS ACOUSTIC CORRELATES

Languages can be labelled according to a rhythm typology proposed by linguists. However, rhythm is complex and some languages do not perfectly match this typology and the search for acoustic correlates has been proposed to evaluate this linguistic classification.

Experiments reported here focus on 5 European languages (English, French, German, Italian and Spanish) and 2 Asian languages (Mandarin and Japanese). According to the linguistic literature, French, Spanish and Italian are syllable-timed languages while English and German are stress-timed languages. Regarding Mandarin, classification is not definitive but recent works tend to affirm that it is a stress-time language (Komatsu et al., 2004). The case of Japanese is different since it is the prototype of a third rhythmic class, namely the mora-timed languages for which timing is related to the frequency of morae[4]. These three categories are related to the notion of isochrony and they emerged from the theory of rhythm classes introduced by Pike, developed by Abercrombie, (1967) and enhanced with mora-timed class by Ladefoged, (1975). More recent works, based on the measurement of the duration of inter-stress intervals in both stress-timed and syllable-timed languages provide an alternative framework in which these

[4] Morae can consist of a V, CV or C. For instance, [kakemono] (scroll) and [nippoŋ] (Japan) must both be divided in four morae: [ka ke mo no] and [ni p po ŋ] (Ladefoged, 1975, p.224).

discrete categories are replaced by a continuum (Dauer, 1983) where rhythmic differences among languages are mostly related to their syllable structure and the presence (or absence) of vowel reduction.

The syllable structure is closely related to the phonotactics and to the accentuation strategy of the language. While some languages will allow only simple syllabic patterns (CV or CVC), other will permit much more complex structures for the onset, the coda or both (e.g. syllables with up to 6 consonants in the coda[5] are encountered in German). Table 1, adapted from Greenberg (1998) displays a comparison of the syllabic forms from spontaneous speech corpora in Japanese and American English.

TABLE 1

The most striking statement is that in both languages, the CV and CVC forms stand for nearly 70% of the encountered syllables. However, the other forms reveal significant differences in the syllabic structure. On the one hand, consonantal clusters are rather common in American English (11.7% of the syllables) while they are almost absent from the Japanese corpus. On the other hand, VV transitions are present in 14.8% of the Japanese syllables while they could only occur by resyllabification at word boundaries in English. These observations roughly correspond with our knowledge of the phonological structure of the words in those two languages. However, the nature of the corpora (spontaneous speech) widely influences the relative distribution of each structure. With read speech (narrative texts), Delattre and Olson (1969) found fairly different patterns for British English: CVC (30.1%), CV (29.7%), VC (12.6%), V (7.4%) and CVCC (7%). CCV that occurs 5.1% in the Switchboard corpus represents only 0.49% of the syllables in the Delattre and Olson corpus. However, statistics

[5] For instance, “you shrink it” will be translated *du schrumpfst’s* [duː ʃrʊmpfsts]. This example is taken from (Möbius, 1998).

calculated on the Switchboard corpus show that 5000 different syllables are necessary to cover 95% of the vocabulary[6] (Greenberg, 1997) and thus that inter-language differences are not restricted to high-frequency syllabic structures. These broad phonotactic differences explain at least partially the mora-time vs. stress-time opposition. Still, studying the temporal properties of languages is necessary to determine whether the rhythm is totally characterized by syllable structures or not.

Beyond the debate on the existence of rhythmic classes (opposed to a rhythmic continuum), the measurement of the acoustic correlates of rhythm is essential for automatic language identification systems based on rhythm. The first statistics made by Ramus, Nespor and Mehler with an *ad hoc* multilingual corpus of 8 languages led to a renewal of interest for these studies (Ramus et al, 1999). Following Dauer, they searched for duration measurements that could be correlated with vowel reduction (resulting in a wide range of duration for vowels) and with the syllable structure. They came up with two reliable parameters: 1. the percentage of vocalic duration%V and 2. the standard deviation of the duration of the consonant intervals ΔC both estimated over a whole utterance. They provided a 2-dimension space in which languages are clustered according to their rhythm class[7]. These results are very promising and prove that in nearly ideal conditions (manual labelling, homogeneous speech rates, etc.), it is possible to find acoustic parameters that cluster languages into explainable categories. The extension of this approach to ALI necessitates the evaluation of these parameters with more languages and less constrained conditions. This raises several problems that can be summarized as follows:

---

[6] This number falls to 2000 syllables necessary to cover 95% of the corpus (i.e. taking into account the frequency of occurrence of each word of the vocabulary).

[7] Actually, the clustering seems to be maximum along one dimension derived from a linear combinaison of ΔC and%V.

- Adding speakers and languages will add inter-speaker variability. Would it result in an overlap of the language-specific distributions?

- Which part of the duration variation observed in rhythmic unit is due to language-specific rhythm and which part is related to speaker-specific speech rate?

- Is it possible to take these acoustic correlates into account for ALI?

A recent study (Grabe & Low, 2000) answers partially to the first question. Considering 18 languages and relaxing constraints on the speech rate, Grabe and Low have found that the studied languages spread widely without visible clustering effect in a 2-dimension space somewhat related to the%V/ΔC space. However, in their study, each language is represented by only 1 speaker, which prevents from drawing definite conclusion on the discrete or continuous nature of the rhythm space. Addressing the variability issue between speakers, dialects and languages, similar experiments focusing on dialects are in progress in our group (Hamdi et al., 2004; Ferragne & Pellegrino, 2004). Though it is beyond the scope of this paper, the second question is essential. Speech rate involves computing a number of certain *units* per second; choosing the appropriate unit(s) remains controversial (syllable, phoneme or morpheme) and so is the interpretation of the measured rate: few units per second means long units, but does it mean that the units are intrinsically long or is the speaker an especially slow speaker? Moreover, the *variation* of speech rate within an utterance is also relevant: the speaking rate of a hesitating speaker may switch from local high values to very low values during disfluencies (silent or filled pauses, etc.) along a single utterance. Consequently, fast variations may be masked according to the time span used for the estimation and the overall speech rate estimation may not be relevant. Besides, the estimation of speech rate is also relevant for automatic speech recognition, since recognizers' performances usually decrease when they come to dealing with especially fast or slow speakers (Mirghafori et al., 1995). For this reason, algorithms exist to estimate either phone rate or syllable rate (e.g. Verhasselt & Martens, 1996; Pfau & Ruske,

1998). However, the subsequent normalization is always applied in a monolingual context, and no risk of masking language specific variation can occur. At present, the effect of this kind of normalization in a multilingual framework has not been studied extensively though it will be essential for ALI purposes. Our group has elsewhere addressed this issue in a study of inter-languages differences of speech rate in terms either of syllables per second or phonemes per second (Pellegrino et al., 2004; Rouas et al., 2004).

The last question is the main issue addressed in this paper. This section assesses the existence of acoustic correlates of the linguistic rhythmic structure. However, whether they are detectable and reliable enough to perform ALI or not is to be tackle. The following sections thoroughly focus on this issue.

4. RHYTHM MODELLING FOR ALI

4.1. Overview of related works

The controversies about the status of rhythm illustrate the difficulty to segment speech into meaningful rhythmic units and emphasize that a global multilingual model of rhythm is a long range challenge. As a matter of fact, even if correlates between speech signal and linguistic rhythm exist, developing a relevant representation of it and selecting an appropriate modelling paradigm is still at stakes.

Among others, Thymé-Gobbel & Hutchings, (1999) have emphasized the importance of rhythmic information in language identification systems. They developed a system based on likelihood ratio computation from the statistical distribution of numerous parameters related to rhythm and based on syllable timing, syllable duration and amplitude (224 parameters are considered). They obtained significant results, and proved that mere prosodic cues can distinguish between some language pairs of the telephone speech OGI-MLTS corpus with results comparable to some non-prosodic systems (depending on the language pairs, correct discrimination rates range from chance to 93%). Cummins and colleagues (1999) have

combined the delta-F0 curve and the first difference of the band-limited amplitude envelope with neural network models. The experiments were also conducted on the OGI-MLTS corpus, using pairwise language discrimination for which they obtained up to 70% of correct identification. The conclusions were that F0 was a more effective discriminant variable than the amplitude envelope modulation and that discrimination is better across prosodic family languages than in the same family.

Ramus and colleagues have proposed several studies (Ramus et al., 1999; Ramus & Mehler, 1999; Ramus, 2002) based on the use of rhythm for language identification. This approach has been furthermore implemented in an automatic modelling task (Dominey & Ramus, 2000). Their experiment aimed at assessing whether an artificial neural network may extract rhythm characteristics from sentences *manually labelled* in terms of consonants and vowels or not. Using the RMN "Ramus, Nespor, Mehler" corpus (1999), they reached significant discrimination results between languages belonging to different rhythm categories (78% for English/Japanese pair) and chance level for languages belonging to the same rhythm category. They concluded that those Consonant/Vowel sequences carry a significant part of the rhythmic patterns of the languages and that they can be modelled. Interestingly, Galves and colleagues (Galves et al., 2002) have reached similar results with no need for hand labelling: Using the RMN data, they automatically derived two criteria from a sonority factor. These two criteria (the mean value $\overline{S}$ and the mean value of the derivate δS of the sonority factor S) lead to a clustering of the languages closely related to the one obtained by Ramus and colleagues. Moreover, δS exhibits a linear correlation with ΔC and $\overline{S}$ is correlated to%V, tending to prove the consistency between the two approaches.

This quick overview of the rhythmic approaches to automatic language identification shows that several approaches, directly exploiting acoustic parameters without explicit unit modelling (e.g. Hidden Markov Model), may significantly discriminate some language pairs. Consequently, rhythm may be relevant for automatic discrimination or identification of the rhythm category of several languages. However, the fact that all these automatic systems exhibit results from

“simple” pairwise discrimination emphasizes that using rhythm in a more complex identification task (with more than two languages) is not straightforward.

4.2. Rhythm unit modelling

The main purpose of this study is to provide an automatic segmentation of the signal into rhythmic units relevant for the identification of languages and to model their temporal properties in an efficient way. To this end, we use an algorithm formerly designed to model vowel systems in a language identification task (Pellegrino & André-Obrecht, 2000). The main features of this system are reviewed hereunder. This model does not pretend to integrate all the complex properties of linguistic rhythm and more specifically, hence it provides by no way a linguistic analysis of the prosodic systems of languages; the temporal properties observed and statistically modelled result from the interaction of several suprasegmental properties and an accurate analysis of this interaction is not yet possible.

Figure 1 displays the synopsis of the system. A language-independent processing parses the signal into vowel and non-vowel segments. Parameters related to the temporal structure of the rhythm units are then computed and language-specific rhythmic models are estimated. During the test phase, the same processing is performed and the most likely language is determined following the Maximum Likelihood rule (see Section 5.2 for more details).

FIGURE 1

In order to extract features related to the potential consonant cluster (number and duration of consonants), a statistical segmentation based on the "Forward-Backward Divergence" algorithm is applied. Interested readers are referred to (André-Obrecht, 1988) for a comprehensive and detailed description of this algorithm. It identifies boundaries corresponding with abrupt changes in the wave spectrum resulting in two main categories of segments: short segments (bursts, but also transient parts of voiced sounds) and longer segments (steady parts of sounds).

A segmental Speech Activity Detection (SAD) is performed to discard long pauses (not related to rhythm), and, finally, the vowel detection algorithm locates sounds matching a vocalic structure via a spectral analysis of the signal. The SAD detects the less intense segment of the utterance (in term of energy) and the others segments are classified as Silence or Speech according to an adaptive threshold; Vowel detection is based on a dynamic spectral analysis of the signal in Mel frequency filters (both algorithms are detailed in Pellegrino & André-Obrecht, 2000). An example of the Vowel/Non-Vowel parsing is provided in Figure 2 (vertical dot lines).

FIGURE 2

TABLE 2

The algorithm is applied in a language- and speaker-independent way without any manual adaptation phase. It is evaluated with the Vowel Error Rate metric (*VER*) defined as follows:

$$VER = 100.\left(\frac{Ndel + Nins}{Nvow}\right)\% \qquad (1)$$

where *Ndel* and *Nins* are respectively the number of deleted vowels and inserted vowels, and *Nvow* is the actual number of vowels in the corpus.

Table 2 displays the performance of the algorithm for spontaneous speech, compared to other systems. The average value reached on 5 languages (22.9% of VER) is as good as the best systems optimized for a given language. The algorithm may be expected to perform better with read speech. However, no phonetically hand-labelled multilingual corpus of read speech was available to the authors to confirm this assumption.

The processing provides a segmentation of the speech signal in pause, non-vowel and vowel segments (see Figure 2). Due to the intrinsic properties of the algorithm (and especially the fact that transient and steady parts of a phoneme may be separated), it is somewhat incorrect to consider that this segmentation is exactly a Consonant/Vowel segmentation since by nature, segments are shorter than phonemes. More specifically, vowel duration is on average

underestimated since attacks and damping are often segmented as transient segments. Figure 2 displays also examples of over-segmentation problems with consonants: the final /fnə/ sequence is segmented into 8 segments (4 for the consonantal cluster, 1 for the vowel steady part and 3 for the final damping). However, our hypothesis is that this sequence is significantly correlated to the rhythmic structure of the speech sound; and the correlation already mentioned between actual syllabic rhythm and its estimation using vowel detection (Pellegrino et al., 2004) confirms this. Our assumption is that this correlation enables a statistical model to discriminate languages according to their rhythmic structure.

Even if the optimal rhythmic units may be language-specific (syllable, mora, etc.), the syllable may be considered as a good compromise. However, the segmentation of speech into syllables seems to be a language-specific mechanism even if universal rules related to sonority and if acoustic correlates of the syllable boundaries exist (see Content, 2000). Thus no language-independent algorithm can be derived at this moment, and even language-specific algorithms are uncommon (Kopecek, 1999; Shastri et al., 1999).

For these reasons, we introduce the notion of Pseudo-Syllables (PS) derived from the most frequent syllable structure in the world, namely the CV structure (Vallée et al., 2000). Using the vowel segments as milestones, the speech signal is parsed into patterns matching the structure: .C^nV. (with *n* an integer that may be zero).

For example, the parsing of the sentence displayed in Figure 2 results in the following sequence of 11 pseudo-syllables:

(CCV.CV.CV.CCCV.CCCV.CCV.CV.CCCV.CCCCV.CCCCV.CCCCCV)

roughly corresponding to the following phonetic segmentation :

(aɪ.hæ.və.p^h.ɻp.blə.mwɪ.ðmaɪ.wɔː.tə.sɒ.fnə)

As said before, the segments labelled in the PS sequence are shorter than phonemes; consequently the length of the consonantal cluster is to a large extent biased to higher values than those given by a phonemic segmentation. We are aware of the limits of such a basic rhythmic parsing, but it provides an attempt to model rhythm that may be subsequently improved. However, it has the considerable advantage that neither hand-labelled data nor extensive knowledge of the language rhythmic structure is required.

A pseudo-syllable is described as a sequence of segments characterized by their duration and their binary category (Consonant or Vowel). This way, each pseudo-syllable is described by a variable length matrix. For example, a .CCV. pseudo-syllable will give:

$$P_{.CCV.} = \begin{bmatrix} C & C & V \\ d_{C1} & d_{C2} & d_{V1} \end{bmatrix} \tag{2}$$

where C and V are binary labels and d_X is the duration of the segment X.

This variable length description is the most accurate, but it is not appropriate for Gaussian Mixture Modelling (GMM). For this reason, another description resulting in a constant length description for each pseudo-syllable has been derived. For each pseudo-syllable, three parameters are computed, corresponding respectively with total consonant cluster duration, total vowel duration and complexity of the consonantal cluster. With the same .CCV. example, the description becomes:

$$P'_{CCV.} = \{(d_{C1} + d_{c2}) \quad d_V \quad N_C\} \tag{3}$$

where Nc is the number of segments in the consonantal cluster (here, $Nc = 2$).

Even if this description is clearly not optimal since the individual information on the consonant segments is lost, it takes a part of the complexity of the consonant cluster into account.

5. LANGUAGE IDENTIFICATION TASK

5.1. Corpus Description and Statistics

Experiments are performed on the MULTEXT multilingual corpus (Campione & Véronis, 1998), extended with Japanese (Kitazawa, 2002) and Mandarin (Komatsu et al., 2004). This database thus contains recordings of 7 languages (French, English, Italian, German, Japanese, Mandarin and Spanish), pronounced by 70 different speakers (5 male and 5 female per language).

The MULTEXT data consist of read passages that may be pronounced by several speakers. Despite the relative small amount of data and to avoid possible text dependency, the following experiments are performed with 2 subsets of the corpus defining no-overlapping training and test sets in terms of speakers and texts (see Table 3). The training corpus is supposed to be representative of each language syllabic inventory. For instance, the mean duration of each passage for the French data is 98 syllables (± 20 syllables) and the overall number of syllable tokens in the French corpus is about 11 700[8]. Even if the syllable inventory is not exhaustive in this corpus, it is reasonable to assume that a statistical model derived from these data will be statistically representative of most of the syllable diversity for each language.

In the classical rhythm typology, French, Italian and Spanish are known as syllable-timed languages while English, German and Mandarin are stress-timed. Japanese is the only mora-timed language of the corpus. Whether this typology is correct or results from an artefact of a rhythmic continuum, our approach should be able to capture features linked to the rhythm structure of these languages.

---

[8] This number takes the number of repetitions of each passage into account. Considering each passage once, the number of syllables is 3 900.

TABLE 3

Intuitively, the duration of consonantal clusters is supposed to be correlated to the number of segments constituting the cluster. Table 4 gives the results of a linear regression with *Dc* (in seconds) as a predictor of *Nc*. For each language, a significant positive correlation is achieved and R^2 values range from 0.71 for French to 0.77 for English and German (see Figure 3 for the scatter plot of English data). In term of slope, values range from 0.0271 for Mandarin to 0.0362 for Spanish meaning that the relation between *Nc* and *Dc* is to some extend language dependent. For this reason, both parameters have been taken into account in the following experiments.

TABLE 4

FIGURE 3

In order to test hypotheses on language specific differences in the distribution of the parameters, a Jarque-Bera test of normality was performed. It confirms that the distributions are clearly non normal ($p<.0001$; $j > 10^3$ for *Dc*, *Dv* and *Nc*, for all languages). Consequently, a non parametric Kruskal-Wallis test was performed for each parameter to evaluate the differences among the languages. They reveal a highly significant global effect of the language for *Dv* ($p<.0001$; df = 6; chi-square = 2248), *Dc* ($p<.0001$; df = 6; chi-square = 1061) and *Nc* ($p<.0001$; df = 6; chi-square = 2839). The results of the Kruskal-Wallis test have then been used in a multiple comparison procedure using Tukey criterion of significant difference.

TABLE 5

TABLE 6

TABLE 7

Table 5 to 7 gives the results of the pairwise comparison. In order to make the interpretation easier, a graphical representation is drawn from the values (Figure 4). Regarding consonant duration, a cluster grouping the stress-timed languages is clearly identified. This cluster is coherent with the complex onsets and coda present in theses languages, either in number of phonemes (English and German) or intrinsic complexity of the consonants (aspirated, retroflex, etc. for Mandarin) The other languages spread along the *Dc* dimension and Japanese and Italian are intermediate between the most prototypical syllable-timed languages (Spanish and French) and the stress-timed languages cluster.

The situation revealed by *Dv* is quite different: English, Japanese, German and Italian cluster together (though significant differences exist between Italian on one side, and English, Japanese and German on the other side) while Mandarin and French are distant. Spanish is also individualized at this opposite extreme of this dimension. *Nc* distributions exhibit important diversity among languages since English and Mandarin are the only cluster for which no significant difference is observed.

FIGURE 4

5.2. GMM modelling for identification

GMM (Gaussian Mixture Models) are used to model the pseudo-syllables which are represented in the three-dimensional space described in the previous section. They are estimated using the EM (Expectation-Maximization) algorithm initialized with the LBG algorithm (Reynolds, 1995; Linde et al., 1980).

Let $X = \{x_1, x_2, \ldots, x_N\}$ be the training set and $\Pi = \{(\alpha_i, \mu_i, \Sigma_i),\ 1 \le i \le Q\}$ the parameter set that defines a mixture of Q p-dimensional Gaussian pdfs. The model that maximizes the overall likelihood of the data is given by:

$$\Pi^* = \arg\max_{\Pi} \prod_{i=1}^{N} \left\{ \sum_{k=1}^{Q} \frac{\alpha_k}{(2\pi)^{p/2}\sqrt{|\Sigma_k|}} \exp\left[-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right] \right\} \qquad (4)$$

where α_k is the mixing weight of the k^{th} Gaussian term.

The maximum likelihood parameters Π^* are obtained using the EM algorithm. This algorithm presupposes that the number of components Q and the initial values are given for each Gaussian pdf. Since these values greatly affect the performances of the EM algorithm, a Vector Quantization (VQ) is applied to the training corpus to optimize them.

The LBG algorithm (Linde et al., 1980) is applied to provide roots for the EM algorithm. it performs an iterated clustering of the learning data into codewords optimized according to the nearest neighbor rule. The splitting procedure may be stopped either when the variation of the data distortion drops under a given threshold or when a given number of codewords is reached (this option is used here).

During the identification phase, all the PS detected in the test utterance are gathered and parameterized. The likelihood of this set of segments $Y = \{y_1, y_2, \ldots, y_N\}$ according to each VSM (denoted L_i) is given by:

$$\Pr(Y|L_i) = \sum_{j=1}^{N} \Pr(y_j|L_i) \qquad (5)$$

where $\Pr(y_j|L_i)$ denotes the likelihood of each segment that is given by:

$$\Pr(y_j|L_i) = \sum_{k=1}^{Q_i} \frac{\alpha_k^i}{(2\pi)^{p/2}\sqrt{|\Sigma_k^i|}} \exp\left[-\frac{1}{2}(y_j - \mu_k^i)^T \Sigma_k^{-1}(y_j - \mu_k^i)\right] \qquad (6)$$

Furthermore, hypothesizing under the *Winner Takes All* (WTA) assumption (Nowlan, 1991), the expression (7) is then approximated by:

$$\Pr(y_j|L_i) = \max_{1\le k\le Q_i}\left[\frac{\alpha_k^i}{(2\pi)^{p/2}\sqrt{|\Sigma_k^i|}} \exp\left[-\frac{1}{2}(y_j - \mu_k^i)^T \Sigma_k^{-1}(y_j - \mu_k^i)\right]\right] \qquad (7)$$

5.3. Automatic Identification results

Pseudo-syllable segmentation has been conceived to be related to language rhythm. In order to assess whether this is actually verified or not, a first experiment aiming at discriminating between the three rhythmic classes is performed; a language identification experiment with the 7 languages is then achieved. At last, a standard acoustic approach is implemented and tested with the same task to provide a comparison.

The first experiment aims at identifying to which rhythmic group belongs the language spoken by an unknown speaker of the MULTEXT corpus. The stress-timed language group gather English, German and Mandarin. French, Italian and Spanish define the syllable-timed language group. The mora-timed language group consists only of Japanese. The number of Gaussian components is fixed to 16 using the training set as a development set to optimize the number of Gaussian components of the GMM. The overall results are presented in Table 8 in a confusion matrix. 119 from 139 files of the test set are correctly identified. The mean identification rate is 86 ± 6% of correct identification (chance level is 33%) and scores range from 80% for syllable- and mora-timed languages to 92% for stress-timed languages. These first results show that the PS approach is able to model temporal features that are relevant for rhythmic group identification.

TABLE 8

The second experiment aims at identifying which of the 7 languages is spoken by an unknown speaker of the MULTEXT corpus. The number of Gaussian components is fixed to 8, using the training set as a development set to optimize the number of Gaussian components of the GMM. The overall results are presented in Table 9 in a confusion matrix. 93 from the 139 files of the test set are correctly identified. The mean identification score thus reaches 67 ± 8% of correct

identification (chance level is 14%). Since the test corpus is very limited, the confidence interval is pretty wide.

TABLE 9

Scores broadly vary and range from 30% for Spanish to 100% for French. Actually, Spanish is massively confused with French; Italian is also fairly misclassified (55% of correct decision) and especially with English. Bad classification is also observed for Mandarin which is confused with both German and English (55% of correct identification). It thus tends to confirm that the classification of Mandarin as a stress-timed language is consistent with the acoustic measurements performed here and for which the Mandarin PS distributions are not significantly different from either German or English distributions.

The wide range of variation observed for the scores may be partially explained studying the speaking rate variability. As for rhythm, speaking or speaker rate is difficult to define but it may be evaluated in term of syllable or phoneme per second. Counting the number of vowels detected per second may provide a first approximation of the speaking rate (see Pellegrino et al., 2004, for a discussion about the speaking rate measurement). Table 10 displays for each language of the database the mean and standard deviation of the number of vowels detected per second among the speakers of the database.

TABLE 10

This rate ranges from 5.05 for Mandarin to 6.94 for Spanish and these variations may be due to both socio-linguistic factors and rhythmic factors related to the structure of the syllable in those languages. Spanish and Italian exhibit the greatest standard deviations (resp. 0.59 and 0.64) of their rate. It means that their models are probably less robust than the others since the parameter distributions are wider. On the opposite, French dispersion is the smallest (0.33) and consistently has the better language identification rate. This hypothesis is supported by a correlation test (Spearman rank order estimation) between the language identification score and

speaking rate standard deviation (ρ = - 0.77, p = 0.05). This shortcoming points out that, at this moment, no normalization is performed on the *Dc* and *Dv* durations. This limitation prevents our model from being adapted to spontaneous speech and this major bottleneck must be tackled in a near future.

At last, the same data and task have been used with an acoustic GMM classifier in order to compare the results of the purely rhythmic approach proposed in this paper with those obtained with a standard approach. The parameters are computed on each segment issued from the automatic segmentation (Section 4). The features consist of 8 Mel Frequency Cepstral Coefficients, their derivatives, and energy, computed on each segment. The number of Gaussian components is fixed to 16 using the training set as a development set to optimize the number of Gaussian components of the GMM. Increasing the number of components does not result in better performances; this may be due to the limited size of the training set both in terms of duration and number of speakers (only 8 speakers per language, except for Japanese: 4 speakers). The overall results are presented in Table 11 in a confusion matrix. 122 from 139 files of the test set are correctly identified. The mean identification rate is 88 ± 5% of correct identification.

TABLE 11

German, Mandarin and Japanese are perfectly identified. The worst results are reached for Italian (65%). Noteworthy is that Mandarin is well discriminated from English and German, contrary to what was observed with rhythmic models. This suggests that the two approaches may be efficiently combined to improve the performances. However, the fact that the acoustic approach reaches significantly better results than the rhythmic approach implies that further improvement are necessary before designing an efficient merging architecture.

6. CONCLUSION AND PERSPECTIVES

While most of the systems developed nowadays for language identification purposes are based on phonetic and/or phonotactic features, we believe that using other kinds of information may be complementary and widen the field of interest of these systems, for example by tackling linguistic typological or cognitive issues about language processing. We propose one of the first approaches dedicated to language identification based on *rhythm* modelling that is tested on a task more complex than pairwise discrimination. Our system makes use of an automatic segmentation into vowel and non-vowel segments leading to a parsing of the speech signal into pseudo-syllabic patterns. Statistical tests performed on the language-specific distributions of the pseudo-syllable parameters show that significant differences exist among the seven languages of this study (English, French, German, Italian, Japanese, Mandarin and Spanish). A first assessment of the validity of this approach is given by the results of a rhythmic class identification task: The system reaches 86 ± 6% of correct discrimination when three statistical models are trained with data from stress-timed languages (English, German and Mandarin), from syllable-timed languages (French, Italian and Spanish) and from Japanese (the only mora-timed language of this study). This experiment shows that the traditional stress-timed vs. syllable-timed vs. mora-timed opposition is assessed with the seven languages we have tested, or more precisely, that the three language groups (English + German + Mandarin vs. French + Italian + Spanish vs. Japanese) exhibit significant differences according to the temporal parameters we propose.

A second experiment done with the 7-language identification task produces relatively good results (67 ± 8% correct identification rate for 21-second utterances). Once again, confusions occur more frequently *within* rhythmic classes than *across* rhythmic classes. Among the seven languages, three are identified with high scores (more than 80%) and can be qualified as “prototypical” from the rhythmic groups (English for stress-timing, French for syllable-timing and Japanese for mora-timing). It is thus interesting to point out that the pseudo-syllable

modelling may also manage to identify languages that belong to the same rhythmic family (e.g. French and Italian are not confused), showing that the temporal structure of the pseudo-syllables is quite language-specific. To summarize, even if the pseudo-syllable segmentation is rough and not able to take the language-specific syllable structures into consideration, it captures at least a part of the rhythmic structure of each language.

However, rhythm can not be reduced to a raw temporal sequence of consonants and vowels, and, as pointed out by Zellner-Keller (2002) its multilayer nature should be taken into account to correctly characterize languages. Among many parameters, those linked to tones or to the stress phenomenon may be pretty salient. For instance, Mandarin, which is fairly confused with other languages in the present study may be well recognized with other suprasegmental features due to its tonal system. Consequently, taking energy or pitch features into account may lead to significant improvement in the language identification performance. However, these physical characteristics lay at the interface between segmental and supra-segmental levels and their values and variations thus result from a complex interaction, increasingly complicating their correct handling.

Besides, the algorithm of pseudo-syllable segmentation may also be enhanced. An additional distinction between voiced and voiceless consonants may be performed to add another rhythmic parameter, and moreover, more complex pseudo-syllables including codas (hence with a C^mVC^n structure) may be obtained by applying segmentation rules based on sonority (see Galves et al., 2002 for a related approach).

Last, the major future challenge will be to tackle the speaking rate variability (shown in Section 5 to be correlated to the identification performance) and to propose an efficient normalizing or modelling that will allow us to adapt this approach to spontaneous speech corpora and to a larger set of languages. Very preliminary experiments performed on the OGI MLTS corpus are reported in Rouas et al., (2003).

7. ACKNOWLEDGEMENTS

The authors would like especially to thank Brigitte Zellner-Keller for her helpful comments and advices and Emmanuel Ferragne for his careful proofreading of the draft of this paper. 188. The authors are very grateful to the reviewers for their constructive suggestions and comments.

This research has been supported by the EMERGENCE program of the Région Rhône-Alpes (2001-2003) and the French Ministère de la Recherche (program ACI “Jeunes Chercheurs” – 2001-2004).

8. REFERENCES

Abercrombie, D., (1967), Elements of General Phonetics, Edinburgh University Press, Edinburgh

Adami, A. G. & Hermansky. H., (2003), Segmentation of Speech for Speaker and Language Recognition, in proc. of Eurospeech, p.841-844, Geneva

André-Obrecht, R., (1988), A New Statistical Approach for Automatic Speech Segmentation, IEEE Trans. on ASSP, vol. 36, n° 1

Antoine, F., Zhu D., Boula de Mareüil P. & Adda-Decker M., (2004), “Approches Segmentales multilingues pour l’identification automatique de la langue : phones et syllabes”, in proc. of *Journées d’Etude de la Parole*, Fes, Morocco

Barkat-Defradas, M., Vasilescu, I., & Pellegrino, F., (2003). Stratégies perceptuelles et identification automatique des langues, Revue PArole

Berg, T. (1992). Productive and perceptual constraints on speech error correction. in: *Psychological Research* 54 pp.114-126.

Besson, M & Schön D, (2001). Comparison between language and music. In "The biological foundations of music" R. Zatorre & I. Peretz, Eds., Ed. Robert J. Zatorre & Isabelle Peretz. Annals of The New York Academy of Sciences, Vol. 930

Bond, Z. S. & Stockmal, V. (2002) Distinguishing samples of spoken Korean from rhythmic and regional competitors. Language Sciences 24, 175-185.

Boysson-Bardies, B., Vihman, M.M., Roug-Hellichius, L., Durand, C., Landberg, I. & Arao, F. (l992) Material evidence of infant selection from the target language: A cross-linguistic study. In C.Ferguson, L. Menn & C. Stoel-Gammon (Eds.), Phonological development: Models, research, implications. Timonium, MD: York Press

Campione, E., & Véronis, J., (1998), A multilingual prosodic database, in Proc. of ICSLP'98, Sydney, Australia

Content, A., Dumay, N., & Frauenfelder, U.H., (2000), The role of syllable structure in lexical segmentation in French, in Proc. of the Workshop on Spoken Word Access Processes, Nijmegen, The Netherlands

Content, A., Kearns, R.K., & Frauenfelder, U.H., (2001), Boundaries versus Onsets in syllabic Segmentation, Journal of Memory and Language, 45(2)

Crystal D. (1990). *A Dictionary of Linguistics and Phonetics*. 3rd Edition. Blackwell Ed. London.

Cummins, F., Gers, F., and Schmidhuber, J., (1999), Language identification from prosody without explicit features, in Proc. of EUROSPEECH '99

Cutler, A., & Norris, D., (1988), The role of strong syllables in segmentation for lexical access, Journal of Experimental Psychology : Human Perception and Performance, 14

Cutler, A., (1996), Prosody and the word boundary problem, in Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition, Morgan & Demuth (Eds.), Mahwah, NJ: Lawrence Erlbaum Associates.

Dauer, R. M., (1983), Stress-timing and syllable-timing reanalyzed, Journal of Phonetics, 11

Delattre, P. & Olsen, C, (1969) "Syllabic Features and Phonic Impression in English, German, French and Spanish", *Lingua* 22: 160-175.

Dominey, P. F., & Ramus, F., (2000), Neural Network Processing of Natural Language: I. Sensitivity to Serial, Temporal and Abstract Structure in the Infant, Language and Cognitive Processes, 15(1)

Drullman, R., Festen, J.M., & Plomp, R., (1994), Effect of reducing slow temporal modulation on speech reception, JASA, 95(5)

Duarte, D, Galves, A., Lopes N. & Maronna, R., (2001). The statistical analysis of acoustic correlates of speech rhythm. Paper presented at the *Workshop on Rhythmic patterns, parameter setting and language change*, ZiF, University of Bielefeld

Ferragne, E. & Pellegrino F., (2004), "Rhythm in Read British English: Interdialect Variability", *to appear in proc. of INTERSPEECH/ICSLP 2004*, October 2004 Jeju, Korea

Fromkin, V. (Ed.) (1973). Speech errors as linguistic evidence. The Hague: Mouton Publishers.

Galves, A., Garcia J., Duarte D. & Galves C., (2002), " Sonority as a Basis for Rhythmic Class Discrimination ", in proc. of the Speech Prosody 2002 conference, 11-13 April 2002

Ganapathiraju, A., (1999), "The webpage of the Syllable Based Speech Recognition Group", http://www.clsp.jhu.edu/ws97/syllable/, last visited July 2002.

Gauvain, J.-L., Messaoudi, A. & Schwenk, H (2004), "Language recognition using phone lattices", in *proc. of International Conference on Spoken Language Processing*, Jeju island, Korea, 2004

Grabe, E. & Low, E.L., (2002), Durational Variability in Speech and the Rhythm Class Hypothesis, Papers in Laboratory Phonology 7, Mouton.

Greenberg, S., (1996), Understanding speech understanding - towards a unified theory of speech perception, in Proc. of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception, Keele, England

Greenberg, S., (1997), On the origins of speech intelligibility in the real world, in Proc. of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-à-Mousson, France

Greenberg, S., (1998), "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation", in Proc. of the ESCA Workshop on Modelling Pronunciation Variation for Automatic Speech Recognition, Kekrade, The Netherlands

Greenberg, S., Carvey, H.M. and Hitchcock, L., (2002), The relation of stress accent to pronunciation variation in spontaneous American English discourse, To appear in the Proc. of the 2001 ISCA Workshop Prosody and Speech Processing, Red Bank, NJ, USA

Hamdi R., Barkat-Defradas M., Ferragne E. & Pellegrino F., (2004), "Speech Timing and Rhythmic structure in Arabic dialects: a comparison of two approaches", *to appear in proc. of INTERSPEECH/ICSLP 2004*, October 2004 Jeju, Korea

Jestead, W., Bacon S.P. & Lehman J.R., (1982), Forward masking as a function of frequency, masker level and signal delay, JASA, 74(4)

Keller, E., & Zellner, B. (1997). Output Requirements for a High-Quality Speech Synthesis System: The Case of Disambiguation. *Proceedings of MIDDIM-96, 12-14 August 96* (pp. 300-308)

Kern S., Davis B.L., Koçbas D., Kuntay A. & Zink I., (to appear). "Crosslinguistic "universals" and differences in babbling", in OMLL - Evolution of language and languages, European Science Fondation

Komatsu, M., Arai, T. & Sugawara, T. (2004): "Perceptual discrimination of prosodic types", in proc. of Speech Prosody, p. 725–728, Nara, Japan, 2004

Kopecek, I., (1999), Speech Recognition and Syllable Segments, in Proc. of the Workshop on Text, Speech and Dialogue - TSD'99, Lectures Notes in Artificial Intelligence 1692, Springer-Verlag.

Ladefoged, P. (1975). *A course in phonetics*. New York: Harcourt Brace Jovanovich pp.296

Levelt, W., & Wheeldon, L., (1994), Do speakers have access to a mental syllabary, Cognition, 50

Li K.P., (1994), Automatic Language Identification using Syllabic Spectral Features, in Proc. of IEEE ICASSP'94, Adelaide, Australia

Liberman, A.M. & Mattingly, I.G., (1985), The motor theory of speech perception revised. Cognition, 21

Linde, Y., Buzo A. & R. M. Gray, "An algorithm for vector quantizer", *IEEE Trans. On COM.*, January 1980, vol. 28, (1980)

MacNeilage, P. (1998). The frame/content theory of evolution of speech production. *Brain and Behavioral Sciences*, 21, 499-546.

MacNeilage, P.F. & Davis, B.L. (2000) Evolution of speech: The relation between ontogeny and phylogeny. In J.R. Hurford, C. Knight & M.G. Studdert-Kennedy (Eds.) The evolutionary emergence of language. Cambridge: Cambridge University Press. 146-160.

MacNeilage, P.F., Davis, B.L., Kinney, A. & Matyear, C.L. (2000). The motor core of speech: A comparison of serial organization patterns in infants and languages. Child Development, 71, 153-163.

Martin, A. F. & M. A. Przybocki. (2003). NIST 2003 Language Recognition Evaluation, in proc. of Eurospeech, p.1341-1344, Geneva

Massaro, D.W., (1972), Preperceptual images, processing time and perceptual units in auditory perception, Psychological Review, 79(2)

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J., (1981), The syllable's role in speech segmentation, Journal of Verbal Learning and Verbal Behavior, 20

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz G., (1996), Coping With Linguistic Diversity: The Infant's Viewpoint, in Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition , Morgan & Demuth (Eds.), Mahwah, NJ: Lawrence Erlbaum Associates.

Mirghafori, N., Fosler, E. & Morgan, N., (1995), Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes, in Proc. of Eurospeech'95, Madrid, Spain

Muthusamy, Y. K., Jain, N., & Cole, R.A., (1994), Perceptual benchmarks for automatic language identification, in Proc. of IEEE ICASSP'94, Adelaide, Australia

Nagarajan T. & Murthy H.A., (2004), "Language Identification Using Parallel Syllable-like Unit Recognition", in proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 401–404, Montreal, Canada

Nazzi, T. & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication 41*(1-2), 233-243.

O'Shaugnessy, D., (1987), Speech Communication. Human and Machine, Addison Wesley, Reading, MA, USA

Ohala, J.J. & Gilbert B., (1979), On listeners' ability to identify languages by their prosody, in Problèmes de prosodie, vol. 2, Léon & Rossi (Eds), Hurtubise HMH

Pellegrino, F., J. Farinas & Rouas J-.L., (2004), "Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech", *in proc. of Speech Prosody 2004*, March 2004, Nara, Japan

Pellegrino, F., & André-Obrecht, R., (2000), Automatic language identification: an alternative approach to phonetic modelling, Signal Processing, Volume 80, Issue 7, July 2000, pp. 1231-1244

Pellegrino, F., J. Farinas & André-Obrecht R., (1999), "Comparison of two phonetic approaches to language identification", in proc. of Eurospeech '99, September 1999, Budapest, Hungary

Pfau, T. & Ruske G., (1998), Estimating the speaking rate by vowel detection, in Proc. of IEEE ICASSP'98, Seattle, WA, USA

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition, 2*, 85-115

Ramus, F., & Mehler, J., (1999), Language identification with suprasegmental cues: A study based on speech resynthesis, Journal of the Acoustical Society of America, 105(1)

Ramus, F., Nespor, M., & Mehler, J., (1999), Correlates of linguistic rhythm in the speech signal, Cognition, 73(3)

Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives. In Proc. of Speech Prosody 2002, Aix-en-Provence, France

Reynolds, D.A., (1995). "Speaker identification and verification using Gaussian mixture speaker models", Speech Communication, Vol. 17, Nos 1-2, 08/95, pp. 91-108

Rouas, J-.L., J. Farinas, Pellegrino F. & Régine André-Obrecht. (2003), "Modeling Prosody for Language Identification on Read and Spontaneous Speech", in proc. of *ICASSP'2003*, Hong Kong, China, p. 40-43

(2004), « Evaluation automatique du débit de la parole sur des données multilingues spontanées », *in actes des XXVèmes JEP*, avril 2004, Fès, Maroc

Shastri, L. Chang, S. & Greenberg, S., (1999), Syllable Detection and Segmentation Using Temporal Flow Neural Networks, in Proc. of ICPhS'99, San Francisco, CA, USA

Singer, E., P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell, & D.A. Reynolds, (2003). Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification, in proc. of Eurospeech, p.1345-1348, Geneva

Stockmal, V., D. Muljani, and Bond, Z. S. (1996) Perceptual features of unknown foreign languages as revealed by multi-dimensional scaling. Proc. of ICSLP Philadelphia, 1748-1751.

Stockmal, V. Moates, D., & Bond, Z. S. (2000) Same talker, different language. Applied Psycholinguistics 21, 383-393.

Taylor, P.A., King S., Isard S.D., Wright H. & Kowtko J., (1997), Using Intonation to Constrain Language Models in Speech Recognition, in Proc. of Eurospeech 97, Rhodes, Greece

Thymé-Gobbel, A., & Hutchins, S. E., (1999), Prosodic features in automatic language identification reflect language typology, in Proc. of ICPhS'99, San Francisco, CA, USA

Todd, N. P. & Brown, G. J., (1994), A computational model of prosody perception, in Proc. of ICSLP'94, Yokohama, Japan

Vallée, N., Boë, L.J., Maddieson, I. & Rousset, I., (2000), Des lexiques aux syllabes des langues du monde – Typologies et structures, in Proc. of JEP 2000, Aussois, France

Vasilescu, I ; Pellegrino, F. ; Hombert, J., (2000), "Perceptual Features for the Identification of Romance Languages", in proc. of ICSLP'2000, Beijing

Verhasselt, J.P. & Martens, J-.P., (1996), A Fast and Reliable Rate of Speech Detector, in Proc. of ISCLP'96, Philadelphia, PA, USA

Wu, S-.L., (1998), Incorporating Information From Syllable-length Time Scales into Automatic Speech Recognition, Report TR-98-014 of the International computer Science Institute, Berkeley, CA, USA.

Weissenborn, J. & Höhle B. (Eds), (2001), Approaches to Bootstrapping. Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition, Volume 1, Acquisition and Language Disorders 23, John Benjamins Publishing Company, 299 p.

Zellner Keller, B., (2002). Revisiting the Status of Speech Rhythm. in Bernard Bel & Isabelle Marlien (eds.), 2002. Proceedings of the Speech Prosody 2002 conference, 11-13 April 2002.(pp. 727-730)

Zellner Keller, B. & Keller, E, (2001). Representing Speech Rhythm. in Keller, E., Bailly, G., Monaghan, A., Terken, J. and Huckvale, M. Eds. Improvements in Speech Synthesis.. Chichester: John Wiley.

Zissman, M. A., Berkling, K. M., "Automatic language identification", *Speech Communication*, Vol. 35, no. 1-2, pp. 115-124, 2001

Table 1 – The ten most common syllabic forms and their frequency of occurrence in Japanese and English. Frequencies are computed on two spontaneous speech corpora. Form in bold are encountered in both languages (adapted from Greenberg, 1998).

JAPANESE		**ENGLISH**	
Form	**% of occurrence**	**Form**	**% of occurrence**
CV	60.4	**CV**	47.2
CVC	17.9	**CVC**	22.1
CVV	11.7	**V**	11.2
V	2.9	**CCV**	5.1
CCV	1.7	**VC**	4.8
CVVC	1.3	CVCC	2.9
CCVV	1.3	**CCVC**	2.5
VC	1.2	VCC	0.5
VV	0.5	CCVCC	0.4
CCVC	0.4	CCCV	0.3
Other	0.7	Other	3.0

Table 2 – Comparison of different algorithms of vowel detection. The formula of the vowel error rate (VER) is given in the text of the paper.

REFERENCE	CORPUS	LANGUAGE	VER
Pfitzinger et al., 1996(*)	PhonDatII (read speech)	German	12.9%
	Verbmobil (spontaneous speech)	German	21.0%
Fakotakis et al., 1997	TIMIT (read speech)	English	32.0%
Pfau & Ruske, 1998	Verbmobil (spontaneous speech)	German	22.7%
Howitt, 2000	TIMIT (read speech)	English	29.5%
Pellegrino & André-Obrecht, 1999	OGI MLTS (spontaneous speech)	French	19.5%
		Japanese	16.3%
		Korean	28.5%
		Spanish	19.2%
		Vietnamese	31.1%
		Average	22.9%

(*) In this study, the error rate is estimated according to syllable nuclei and not explicitly vowels.

Table 3 – The MULTEXT Corpus (from Campione & Véronis, 1998)

LANGUAGE	PASSAGES PER SPEAKER	TOTAL DURATION (MIN.)	AVERAGE DURATION PER PASSAGE (S)	TRAINING (MIN.)	TEST (MIN.)
English	15	44	17.6	24	6
French	10	36	21.9	29	7
German	20	73	21.9	29	7
Italian	15	54	21.7	30	7
Mandarin	15	58	20.0	26	11
Japanese	40	124	31	39	6
Spanish	15	52	20.9	27	8

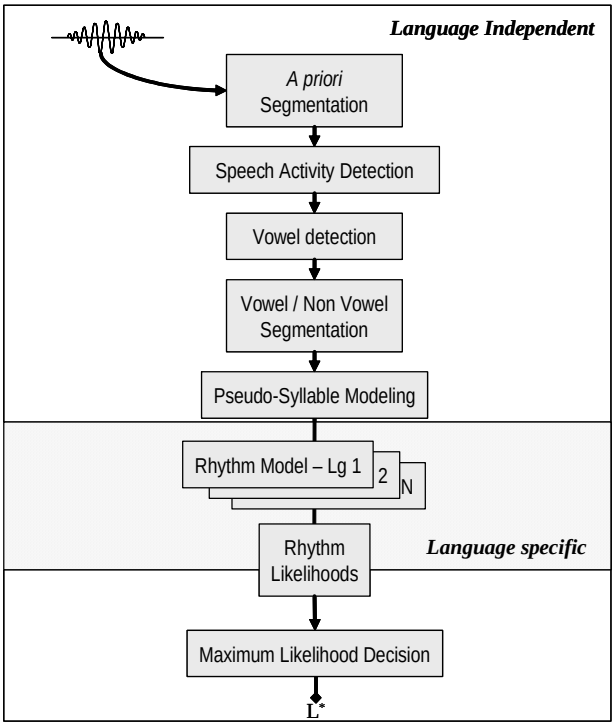

Figure 1 – Synopsis of the implemented system.

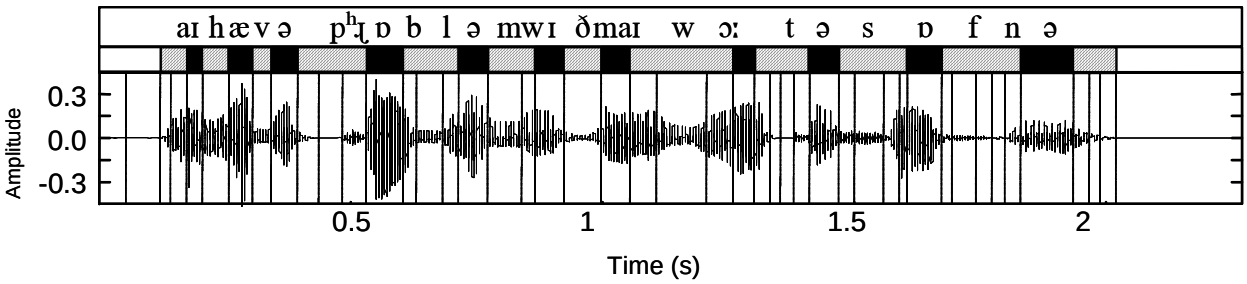

Figure 2 – Example of the automatic vowel/non vowel labelling. The utterance is “I have a problem with my water softener…”. The first tier gives the phonetic transcription. The second tier displays the result of the automatic algorithm (white = pause; dashed = non vowel and black = vowel). Vertical dotted lines displays the result of the *a priori* segmentation.

Table 4 – Estimation of *Dc* as a predictor of *Nc*. Results of a linear regression in least-squares sense. NB PS is the number of pseudo-syllables from which the regression was performed for each language. R^2 is the squared correlation coefficient (according to Spearman rank order estimation). All correlations are highly significant ($p<.0001$).

LANGUAGE	R^2	EQUATION	NB PS
EN	0.83	$100\ \hat{Nc} = 3.68Dc + 22$	11741
FR	0.78	$100\ \hat{Nc} = 3.25Dc + 15$	9307
GE	0.82	$100\ \hat{Nc} = 3.43Dc + 56$	19296
IT	0.81	$100\ \hat{Nc} = 3.27Dc + 34$	14867
JA	0.80	$100\ \hat{Nc} = 3.27Dc + 56$	28913
MA	0.79	$100\ \hat{Nc} = 2.95Dc + 86$	14583
SP	0.80	$100\ \hat{Nc} = 3.76Dc + 20$	15005

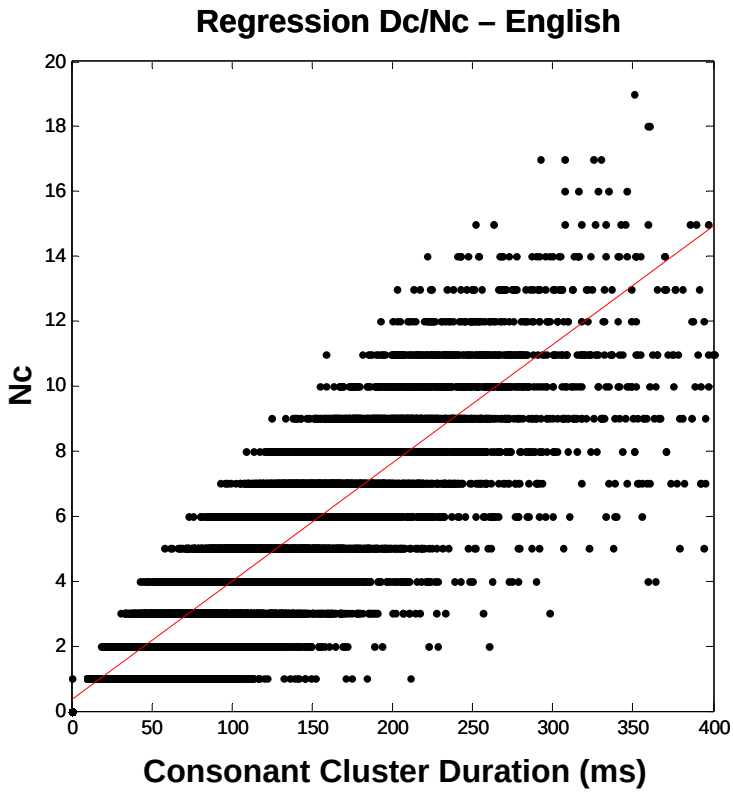

Figure 3 – Evaluation of *Dc* as a predictor of *Nc* for English. Dots are measured values and the solid line is the best linear fit estimated in the least-squares sense.

Table 5 – Significancy of the differences among the distributions of *Dc* (Multiple comparisons from the Kruskal Wallis analysis). *n.s.*is not significant and * is significant or highly significant

	EN	**FR**	**GE**	**IT**	**JA**	**MA**	**SP**
EN		*	*.	*	*	n.s.	*
FR			*	*	*	*	*
GE				*	*	n.s.	*
IT					n.s.	*	*
JA						*	*
MA							*

Table 6 – Significancy of the differences among the distributions of *Dv* (Multiple comparisons from the Kruskal-Wallis analysis). *n.s.*is not significant and * is significant or highly significant

	EN	**FR**	**GE**	**IT**	**JA**	**MA**	**SP**
EN		*.	n.s.	*.	n.s.	*.	*
FR			*	*	*	n.s.	*
GE				*	n.s.	*	*
IT					*	*	*
JA						*	*
MA							n.s.

Table 7 – Significancy of the differences among the distributions of *Nc* (Multiple comparisons from the Kruskal-Wallis analysis). *n.s.*is not significant and * is significant or highly significant

	EN	**FR**	**GE**	**IT**	**JA**	**MA**	**SP**
EN		*	*	*	*	n.s.	*
FR			*	*	*	*	*
GE				*	*	*	*
IT					*	*	*
JA						*	*
MA							*

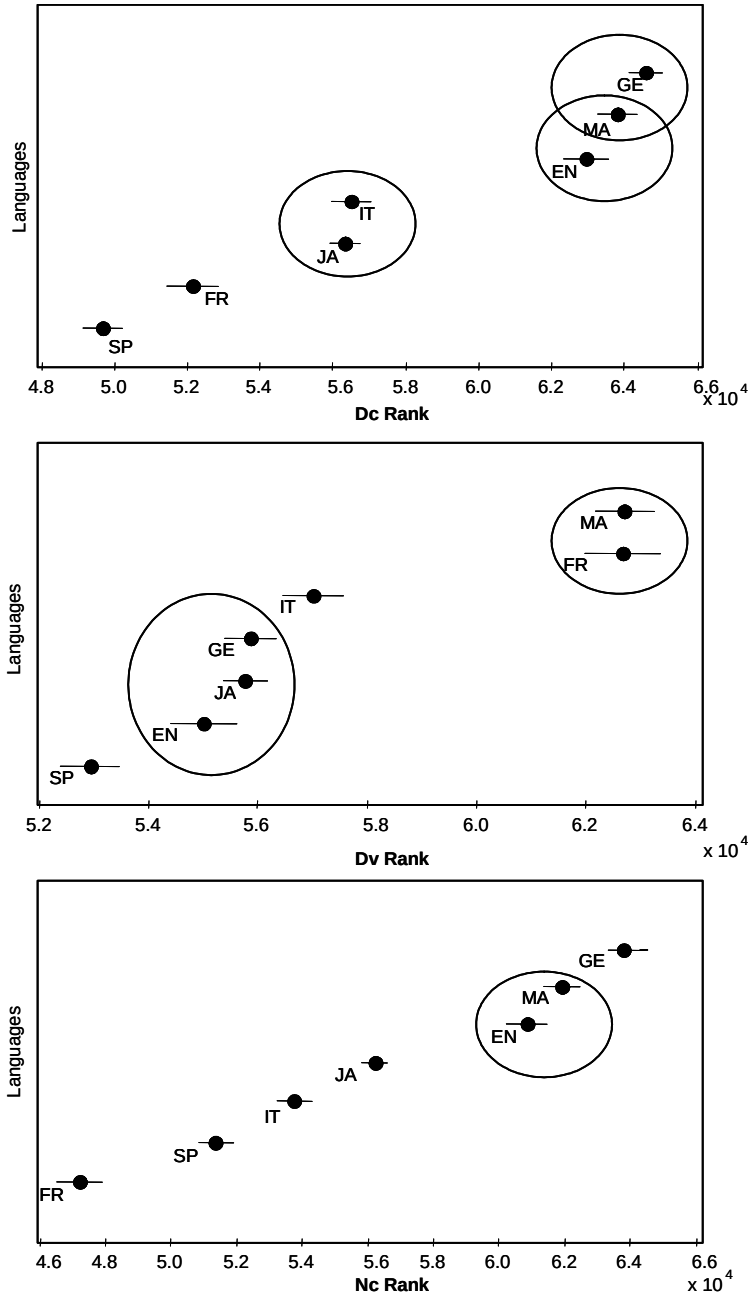

Figure 4 – Estimated rank for each language for the *Dc* distribution above, the *Dv* distribution (middle) and the *Nc* distribution below. Lines spanning across the dots give the 95% confidence interval. Ellipses cluster languages for which the multiple comparisons show no significant differences.

Table 8 – Results for the rhythmic group identification task (16 Gaussian components per GMM). Overall score is 86 ± 6% (119/139 files).

Model / Rhythmic group	Stress-timed	Syllable-timed	Mora-timed
Stress-timed	**55**	5	-
Syllable-timed	10	**48**	1
Mora-timed	2	2	**16**

Table 9 – Results for the 7-language identification task (8 Gaussian components per GMM).

Overall score is 67 ± 8% (93/139 files).

Model / Language	EN	GE	MA	FR	IT	SP	JA
English	**16**	1	1	-	1	1	-
German	5	**14**	1	-	-	-	-
Mandarin	4	3	**11**	-	1	-	1
French	-	-	-	**19**	-	-	-
Italian	6	1	1	-	**11**	-	1
Spanish	-	-	-	8	2	**6**	4
Japanese	2	-	-	-	2	-	**16**

Table 10 – Speaking rate approximated by the number of vowels detected per second for the seven languages.

	English	French	German	Italian	Japanese	Mandarin	Spanish
Mean	5.39	6.37	5.06	5.71	5.29	5.05	6.94
Std Deviation	0.52	0.33	0.45	0.64	0.51	0.52	0.59

Table 11 – Results for the 7-language identification task (standard acoustic approach, 16 Gaussian components per GMM). Overall score is 88 ± 5% (122/139 files).

Model / Language	EN	GE	MA	FR	IT	SP	JA
English	**15**	-	-	-	5	-	-
German	-	**20**	-	-	-	-	-
Mandarin	-	-	**20**	-	-	-	-
French	-	-	-	**17**	-	2	-
Italian	2	-	2	-	**13**	1	2
Spanish	1	-	-	2	-	**17**	-
Japanese	-	-	-	-	-	-	**20**